

MULTI-ARRAY FUSION FOR BEAMFORMING AND LOCALIZATION OF MOVING SPEAKERS

Ilyas Potamitis, George Tremoulis, Nikos Fakotakis, George Kokkinakis

Wire Communications Laboratory, Electrical and Computer Engineering Dept.,
University of Patras, 261 10 Rion, Patras, Greece, Tel:+30 2610 991722, Fax:+30 2610 991855
e-mail: potamitis@wcl.ee.upatras.gr

Abstract

In this work we deal with the fusion of the estimates of independent microphone arrays to produce an improved estimate of the Direction of Arrival (DOA) of one moving speaker, as well as localization coordinates of multiple moving speakers based on Time Delay Of Arrivals (TDOA). Our approach (a) fuses measurements from independent arrays, (b) incorporates kinematic information of speakers' movement by using parallel Kalman filters, and (c) associates observations to specific speakers by using a Probabilistic Data Association (PDA) technique. We demonstrate that a network of arrays combined with statistical fusion techniques provides a consistent and coherent way to reduce uncertainty and ambiguity of measurements. The efficiency of the approach is illustrated on a simulation dealing with beamforming one moving speaker on an extended basis and localization of two closely spaced moving speakers with crossing trajectories.

1. Introduction

The paper suggests ways to combine data from multiple arrays to reduce the uncertainty of the 'state' of the speakers, (where 'state' is the angle and angle rate of arrival for beamforming applications and Cartesian coordinates and velocity for localization techniques). The purpose of employing multiple microphone arrays is to provide observational data of the same speakers from different perspectives. These data are provided in data fusion techniques that integrate the observations from each array to complement the data of other arrays in order to obtain broader coverage of the acoustically sensed environment as well as more accurate speaker-state estimates. Multisensor data fusion has found widespread application in diverse areas ranging from tracking of aircrafts and missiles to navigation [1-3]. In the framework of robust speech acquisition for videoconferencing, multimedia conferencing, speech recognition with regard to distant talkers, etc. data fusion techniques are currently receiving some attention primarily in view of combining observational data from different kind of sensors (e.g. video and audio signals, [4-5]). However, a unifying framework applied to the problem of spatially selective speech acquisition is still missing and most of the theory of multi-target, multi-sensor tracking theory seems to remain unexploited yet. To our point of view, the field of applications of multi-target, multi-sensor theory in robust speech acquisition is wide. To mention two examples of potential applications: in beamforming applications carried out by a single array; in case of two closely spaced speakers who are distant with regard to the array even if the DOA or TDOA estimation techniques manage to give reliable angle

fixes for the wideband sources it is practically impossible to form such a sharp beam that would allow for separation of voices. Again, if two speakers are active while lying across the line of sight of the main reception lobe or when the path between the speaker and the array is obstructed by a very large obstacle, separation of voices is practically infeasible. As regards the localization problem, it has been shown that the reliability of the measurements is highly dependent on the positioning of the arrays [6]. Moreover, tracking either the DOAs or location of multiple moving speakers is severely complicated by the fact that repeated application of DOA estimation or localization techniques over consecutive frames does not yield tracking of speakers [8, 9]. This is because the order in which these estimates appear is highly dependent on the spectral content of each source, preventing successive measurements from being associated with particular speakers. In this work, two fusion schemes that are common in Multisensor Multitarget theory are adapted to the special case of speech signals to integrate observations from multiple microphone arrays. We demonstrate their advantage against single arrays possessing the same total number of microphones. The workhorse of the estimation process is a sequential Kalman-based tracking algorithm that incorporates a model for speakers' motion into the procedure of recursively deriving angle or localization estimates from multiple observations of the same speaker. A data association technique is also incorporated into the state inference scheme that associates unambiguously the observation to speakers and rejects clutter measurements. The present work was supported by the INSPIRE (IST 2001-32746) EC project.

2. Fusion and Tracking

2.1. Interacting Multiple Model estimation

An active speaker's trajectory can be subdivided into distinct segments each corresponding to a different behavioural mode of movement. That is, the speaker can be standing still while talking or walking or performing a turn etc. The multiple Kalman approach (known as Interactive Multiple Model – IMM) assumes that at time instant k the speaker is in one of a finite number of modes. Therefore, speakers' motion can be modeled by two state variables; the first being composed of DOAs or Cartesian coordinates and their rate of change (depending on the application), and a discrete regime variable which describes the distinct segments of motion. Kalman filters operate in parallel; each corresponding to a behavior mode that undergoes jumps from model i to model j , according to a set of transition probabilities modeled with a Markov chain. The problem of IMM state estimation in the context of our work is to infer the kinematic and modal state based on noisy observation of DOAs or coordinates extracted by TDOAs. One cycle of tracking for the IMMs is as follows:

Step 0: An initial estimate of DOA or Cartesian coordinates is provided by DOA or TDOA techniques respectively.

Step 1: Each of the speakers' equation describing their movement and the observation equation is a linear function of the current state. IMM consist of multiple (say j) models. We assume that the speakers' follow each model at time k with probability μ_j . In this work the Newtonian source motion and observational equations become:

$$\mathbf{s}(k)=\mathbf{F}\mathbf{s}(k-1)+\mathbf{w}_j(k) \quad (1)$$

$$\mathbf{y}(k)=\mathbf{H}\mathbf{s}(k)+\mathbf{u}_j(k) \quad (2)$$

$\mathbf{w}_j(k)\sim\mathcal{N}(\mathbf{0}, \mathbf{Q}_j)$ is the Gaussian zero mean process noise vector having covariance matrix \mathbf{Q}_j . $\mathbf{w}_j(k)$ models accelerations experienced by a moving source. $\mathbf{u}_j(k)\sim\mathcal{N}(\mathbf{0}, \mathbf{U}_j)$ is the measurement noise having covariance \mathbf{U}_j . However, not all measurements originate from speakers. False measurements originate mainly due to reverberation and silence periods (DOA techniques return a fixed number of angles regardless of the true number of speakers at a given instant). A validation region for each mode j at time k is constructed around the measurements based on the following predictions:

$$\hat{\mathbf{s}}_j(k|k-1)=\mathbf{F}\hat{\mathbf{s}}_j(k-1|k-1) \quad (3)$$

$$\hat{\mathbf{y}}_j(k|k-1)=\mathbf{H}\hat{\mathbf{s}}_j(k|k-1) \quad (4)$$

$$\mathbf{P}_j(k|k-1)=\mathbf{F}\mathbf{P}_j(k-1|k-1)\mathbf{F}^T+\mathbf{Q}_j \quad (5)$$

Step 2: New measurements are received from the sensors and are validated if they lie inside an acceptance region with probability P_G fulfilling:

$$e_j=(\mathbf{y}(k)-\hat{\mathbf{y}}_j(k|k-1))(\mathbf{S}_j(k))^{-1}(\mathbf{y}(k)-\hat{\mathbf{y}}_j(k|k-1))\leq g_j^2 \quad (5)$$

where $\mathbf{S}_j(k)$ is the covariance of the innovation and g_j^2 (known as the number of standard deviations of the gate) is determined by P_G as well as the dimension of the state from a chi-squared table [9]. The measurement is retained if it is inside the gate and associated to the previous estimates forming a track; otherwise it is rejected in order not to affect the estimation procedure. The PDA filter returns association probabilities (β) for the speaker oriented measurements and (β_0) for the clutter:

$$\beta_j = \frac{e_j}{b + \sum_{i=1}^m e_i}, \quad \beta_0 = \frac{b}{b + \sum_{i=1}^m e_i}, \quad b = \lambda \sqrt{\det(2\pi\mathbf{S})} \frac{1 - P_D P_G}{P_D}$$

λ : density of the clutter, P_D : the detection probability [1], [9]. The Kalman gain is defined for each mode j by:

$$\mathbf{K}_j(k)=\mathbf{P}_j(k|k-1)\mathbf{H}^T(\mathbf{H}\mathbf{P}_j(k|k-1)\mathbf{H}^T+\mathbf{U}_j)^{-1} \quad (6)$$

The a-posteriori state estimate and covariance for mode j are:

$$\hat{\mathbf{s}}_j(k|k)=\hat{\mathbf{s}}_j(k|k-1)+\mathbf{K}_j(k)\sum_j\beta_j(\mathbf{y}(k)-\mathbf{H}\hat{\mathbf{s}}_j(k|k-1)) \quad (7)$$

$$\mathbf{P}_j(k|k)=\beta_0\mathbf{P}_j(k|k-1)+(1-\beta_0)(\mathbf{I}-\mathbf{K}_j(k)\mathbf{H})\mathbf{P}_j(k|k-1)+\mathbf{P}^E \quad (8)$$

$$\mathbf{P}^E=\mathbf{K}_j(k)\sum_j\beta_j(\mathbf{y}(k)-\mathbf{H}\hat{\mathbf{s}}_j(k|k-1))(\mathbf{y}(k)-\mathbf{H}\hat{\mathbf{s}}_j(k|k-1))^T\mathbf{K}_j(k)^T \quad (9)$$

Step 3: The final predicted state and covariances are computed by combining the estimates of all possible modes:

$$\hat{\mathbf{s}}(k|k)=\sum_j\mu_j(k)\hat{\mathbf{s}}_j(k|k) \quad (10)$$

$$\mathbf{P}(k|k)=\sum_j\mu_j(k)[\mathbf{P}_j(k|k)+[\hat{\mathbf{s}}_j(k|k)-\hat{\mathbf{s}}(k|k)][\hat{\mathbf{s}}_j(k|k)-\hat{\mathbf{s}}(k|k)]^T] \quad (11)$$

Due to space limitations, detailed description of IMMs is beyond the scope of this paper (for the update of μ_j and thorough presentation of IMM models see [1] and [9]).

For the beamforming application the DOAs of the two arrays are combined to produce range estimates and the Eq. 1-15 are employed. However, for the localization of multiple speakers a sequential algorithm propagates the estimate of a separate Kalman filter for each sensor i to the other j sensors [7]:

$$\mathbf{s}^i(k|k)=\mathbf{s}(k|k-1)+\mathbf{K}^i(k)\mathbf{v}^i(k) \quad (12)$$

$$\mathbf{s}^i(k|k)=\mathbf{s}^{i-1}(k|k)+\mathbf{K}^i(k)\mathbf{v}^i(k) \quad (13)$$

$$\mathbf{K}^i(k)=\mathbf{P}(k|k-1)\mathbf{H}\mathbf{S}^i(k)^{-1} \quad (14)$$

$$\mathbf{K}^i(k)=\mathbf{P}(k|k,i-1)\mathbf{H}\mathbf{S}^i(k)^{-1} \quad (15)$$

$$\mathbf{P}(k|k,i)=\mathbf{P}(k|k,i-1)\mathbf{H}\mathbf{P}(k|k-1) \quad (16)$$

$$\mathbf{P}(k|k,j)=\mathbf{P}(k|k,i-1)\mathbf{H}\mathbf{P}(k|k,j-1) \quad (17)$$

2.2. Beamforming on a single moving speaker

DOAs for each array are communicated to a central level tracker (IMM filter). The central level tracker combines angle only measurements to form a full position track that includes range, range rate, angle and angle rate. In radar multi-sensor framework this process is denoted as triangulation and its function is illustrated in Fig. 1. Let θ_{1x}, θ_{2x} be the direction cosines of the lines of sight that are computed from the DOA measurements corresponding to Array 1 and Array 2 respectively. The range estimates (R_1, R_2) are obtained by the cosine law [2]. Let $\mathbf{s}(k)$ be the state vector at time k composed by the range $R_1(k)$, range rate $R_1'(k)$, angle θ_1 and the angular velocity θ_1' with respect to the axis of Array 1.

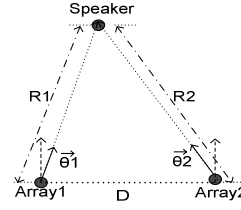


Fig. 1: Triangulation for two microphone arrays.

$$\hat{\mathbf{R}}_1=R_1\boldsymbol{\theta}_1=R_1(\theta_{1x}\mathbf{i}_x+\theta_{1y}\mathbf{i}_y) \quad (12)$$

$$\hat{\mathbf{R}}_2=R_2\boldsymbol{\theta}_2=R_2(\theta_{2x}\mathbf{i}_x+\theta_{2y}\mathbf{i}_y) \quad (13)$$

$$\mathbf{G}=\begin{bmatrix} \theta_{1x} & -\theta_{2x} \\ \theta_{1y} & -\theta_{2y} \end{bmatrix} \quad (14)$$

$$\mathbf{R}=\begin{bmatrix} \hat{\mathbf{R}}_1 \\ \hat{\mathbf{R}}_2 \end{bmatrix}=(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{D}$$

$$\mathbf{s}(k)=[R_1(k) R_1'(k) \theta_1(k) \theta_1'(k)]^T$$

$$\mathbf{y}(k)=[R_1(k) \theta_1(k)]^T$$

Let the variance of range be σ_r^2 and the variance of angle σ_θ^2 . Then $\sigma_{r'}^2=r^6/\cos^2\theta$ and $\sigma_{\theta'}^2=r^2/\cos^2\theta$ (see p. 429 in [10]). We have fixed the design parameters so that the non-moving and slowly moving mode possess low-level process noise ($q=0.001$) while the turning mode (maneuvering model) possesses a much higher noise level ($q=100$). $P_G=0.99$, $P_D=0.99$. DOA estimation is based on the application of wideband MUSIC estimates [10] which are derived from a block of five, half-overlapping 64 ms hamming-windowed frames ($T=160$ msec) using a 512 points FFT at 8 kHz sampling rate. The matrices needed by the Kalman filter are:

$$\mathbf{F}_1=\begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}, \quad \mathbf{Q}=\begin{bmatrix} T^4 & T^3 \\ 4 & 2 \\ T^3 & T^2 \end{bmatrix}, \quad \mathbf{F}_{3a}=\begin{bmatrix} \mathbf{F}_1 & \mathbf{0}_{2,2} \\ \mathbf{0}_{2,2} & \mathbf{F}_1 \end{bmatrix}, \quad \mathbf{Q}_{3a}=\begin{bmatrix} \mathbf{Q} & \mathbf{0}_{2,2} \\ \mathbf{0}_{2,2} & \mathbf{Q} \end{bmatrix}, \quad \mathbf{U}=\begin{bmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_\theta^2 \end{bmatrix}, \quad \mathbf{H}=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

2.3. Localisation of multiple moving speakers

Accurate TDOA estimation techniques under medium reverberation require long data segments to perform some kind of ensemble averaging rendering them unsuitable for moving speakers. The major problem though, with most localization techniques is that they are generally unsuitable for a multi-speaker environment.

In this work we have used a TDOA technique capable of a high update rate and able to distinguish individual speakers in a multipath environment by associating one TDOA per frame to the predominant (in terms of energy) speaker [6, 8]. In this framework, each bilinear array provides a localization estimate per frame (1024 FFT at 20 kHz sampling and 512 samples window with 256 samples overlap). The novelty of our approach compared to [8] is that the two bilinear arrays used to find the TDOAs also provide two independent location estimates that are sent to a central fusion (a IMM model Eq. 1-17) that combines their independent estimates. The corresponding state vector for this case is $\mathbf{s}(k)=[x(k) x'(k) y(k) y'(k) z(k) z'(k)]^T$, $\mathbf{y}(k)=[x(k) y(k) z(k)]^T$ and the corresponding matrices:

$$\mathbf{F}_{6a}=\begin{bmatrix} \mathbf{F}_1 & \mathbf{0}_{2,2} & \mathbf{0}_{2,2} \\ \mathbf{0}_{2,2} & \mathbf{F}_1 & \mathbf{0}_{2,2} \\ \mathbf{0}_{2,2} & \mathbf{0}_{2,2} & \mathbf{F}_1 \end{bmatrix}, \quad \mathbf{Q}_{6a}=\begin{bmatrix} \mathbf{Q} & \mathbf{0}_{2,2} & \mathbf{0}_{2,2} \\ \mathbf{0}_{2,2} & \mathbf{Q} & \mathbf{0}_{2,2} \\ \mathbf{0}_{2,2} & \mathbf{0}_{2,2} & \mathbf{Q} \end{bmatrix}, \quad \mathbf{R}_{6a}=\begin{bmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{bmatrix}, \quad \mathbf{H}_{6a}=\begin{bmatrix} 1 & 0 & 0 & \mathbf{0}_{3,3} \\ 0 & 0 & 1 & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & 0 & 1 & 0 \end{bmatrix}$$

3. Evaluation

As regards the beamforming, the simulation is based on the method of images and for the beamforming application takes place in a typical $6.8\text{m}\times 4.55\text{m}\times 3\text{m}$ room with 30 dB background noise. Each array is composed of $M=8$ omnidirectional microphones with $d=0.1\text{m}$ spacing between microphones and with both arrays located at 1.6m height. The topology of the experiment is designed in a way that analytic derivation of the DOAs is possible and includes a speaking-while-standing mode as well as speaking-while-walking mode. The movement of the speaker is presented in Fig. 2. During his movement the speaker is always active uttering random TIMIT recordings. The speaker starts from (6.4, 0.33, 1.6) meters at $t=0$ and performs a circular movement with angular velocity of 0.1257 rad/sec. At 6.25 seconds stops (at 45°). At $t=8.25$ sec moves on and at 14.5 sec stops moving at 90° with regards to the endfire of the Array 1. At $t=17.5$ sec continues the circular motion until $t=30$ sec. For the localization experiments the first speaker performs the same movement as in the beamforming case but starts from (4.38, 0.68, 1.6)m and ends at (0.62, 0.68, 1.6)m at $t=20$ secs (see Fig. 7).

3.1. Beamforming on a single moving speaker

In Fig. 3 DOA estimation based only on Array 1 (eight microphones with their center located at (3.5, 0.68, 1.6)m) is depicted. In Fig. 4 DOA estimation based only on Array 2 (its center located at (6.5, 0.68, 1.6)m) and in Fig. 5 the results of fusing the outputs of the four central microphones of Array1 and four central microphones of Array 2) are depicted. To summarize the results of the simulation: the combination of the arrays by a central IMM filter returns more reliable DOA estimation in almost all times (see Figs. 3-5). The filter is able to concentrate the DOA estimates when the speaker is still and to smooth DOA estimates while he moves rejecting clutter measurements. The advantage of the geometrically dispersed arrays to obtain better resolution is illustrated in Fig. 5. Large angle estimation errors in the beginning (Fig. 6) are due to the severe resolution problem of Array 2 whose far-field hypothesis is violated and because the speaker is standing in the endfire position of Array 1. However, the combination quickly recovers from the completely erroneous estimates of Array 2 based on the estimates of Array 1. Array2 outperforms the combination in the last seconds because the speaker movement lies at broadside of Array2 and at the endfire position of Array 1. The range estimate degrades due to the erroneous estimate of Array 1 but for most of the time of the movement the cooperation of the arrays is beneficial.

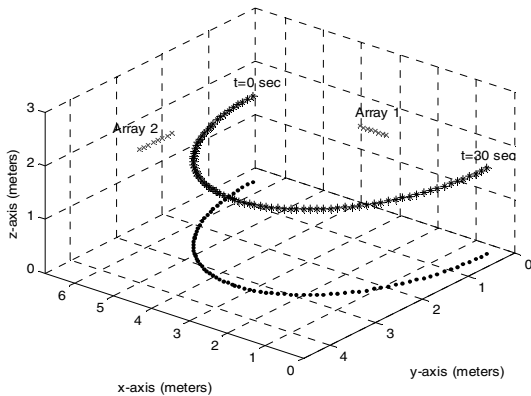


Fig. 2: Trajectory of single moving speaker.

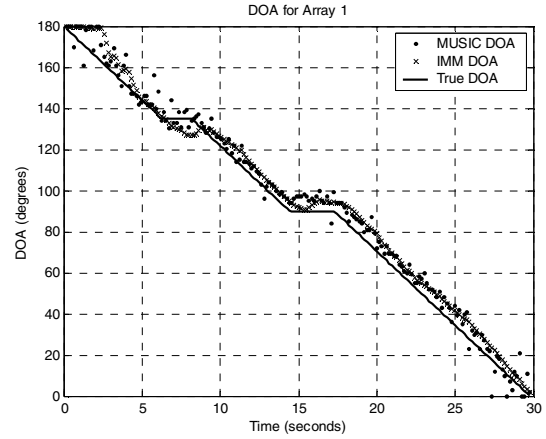


Fig. 3: DOA as seen from Array 1. MUSIC DOA: unprocessed observations from the MUSIC technique, IMM DOA: MUSIC observations processed with IMM-PDA.

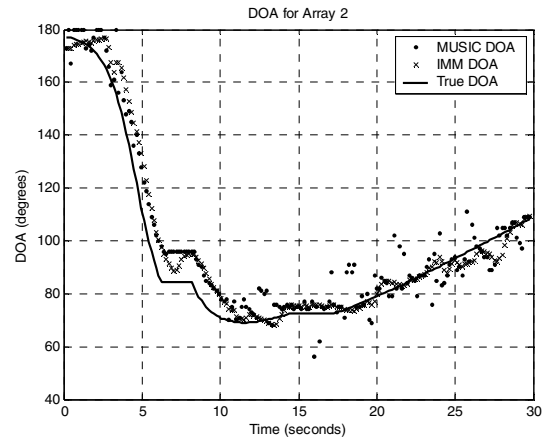


Fig. 4: DOA as seen from Array 2. Notation as in Fig. 3.

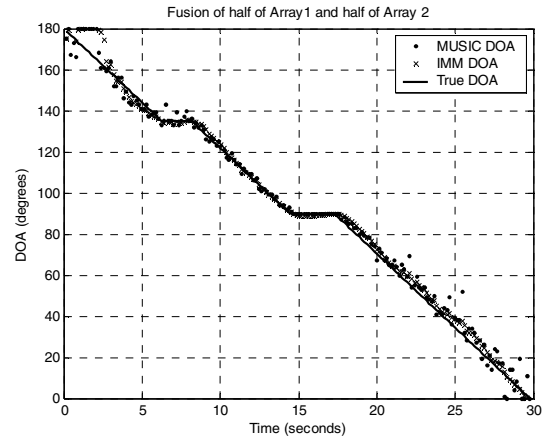


Fig. 5: DOA with respect to the axis of the Array 1 after fusing the observations of Array 1 and Array 2. Note that half of the microphones of each array are used during the fusion.

3.2. Localisation of multiple moving speakers

For the localization application we performed the experiment in a smaller room $5\text{m}\times 3\text{m}\times 3\text{m}$. Two bilinear arrays of 10 microphones with 0.25m spacing are employed (see Fig. 7). At time $t=0$ the second speaker is located at (0.5, 1, 1.6)m and walks for 10 seconds heading parallel to the x-axis at a speed

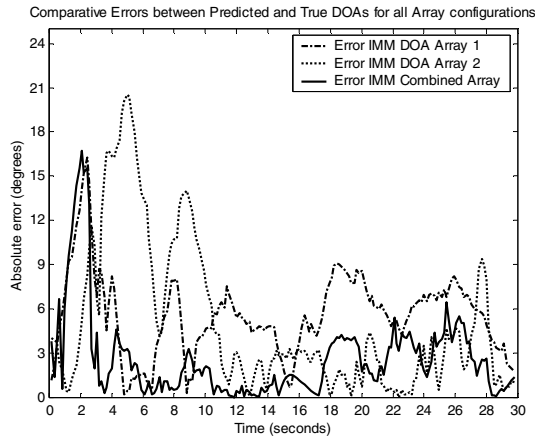


Fig. 6: Absolute DOA error for all array configurations.

of 0.4 m/sec. At (4.5, 1, 1.6)m stops for 3 secs. Then the speaker performs a ninety degree turn and keeps walking parallel to the y axis for 4 sec with a speed 0.25 m/sec. Finally, he stops at (4.5, 2, 1.6)m and talks for 3 secs. The result of the localization experiment is illustrated in Fig. 7. The IMM filters were capable of following the trajectories of the moving speakers and smoothing the location estimates while the gating process of PDA was able to reject clutter measurements and to associate locations with speakers (see Fig. 8). We do not present comparative results with each bilinear array functioning alone, since it completely failed to provide reliable measurements to allow for the tracking of the two speakers. This inability is due to the fact that the TDOA estimation technique described in [6] returns one TDOA per frame of the most dominant speaker which is usually the one closer to the array (for overlapping speech). Therefore the use of a single bilinear array results in sparse location estimates for the speaker far from the array [6]. In our framework the use of distributed subarrays is beneficial both in terms of TDOA estimation and in providing better coverage of the movement of both speakers.

4. Conclusions

A network of arrays is a natural extension of the single array, as an array is an extension of a single microphone. A microphone array provides a spatial selectivity advantage over a single microphone, while a dispersed network of arrays can extract information about the acoustically sensed environment that would be difficult or even impossible for a single array. In large enclosures the array management scheme may not be able to allocate all its processing power to all sensors. The proposed technique paves the way for many more schemes that could reduce computational demands by selecting the proper combinations of arrays or switching among arrays depending on the location of the speaker in the room. In a decentralized approach this would be based on the covariance of the state estimate of each array. Properly tuned Kalman filters and data association techniques ensure that the fusion of information can be done in a way that uncertainty of measurements is reduced. Although sound localization and beamforming was achieved by employing only the acoustic modality, data fusion in the IMM framework can be based on integrating observations from a variety of sensors like infrared and vision [2]. We suggest that Multitarget Multisensor theory can serve as a unifying framework to assist spatial sound selectivity and localization based on microphone.

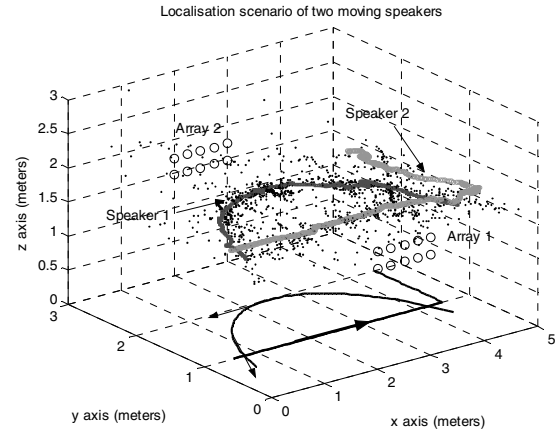


Fig. 7: Estimation of the Cartesian coordinates of two moving speakers by combining the estimation of the two arrays.

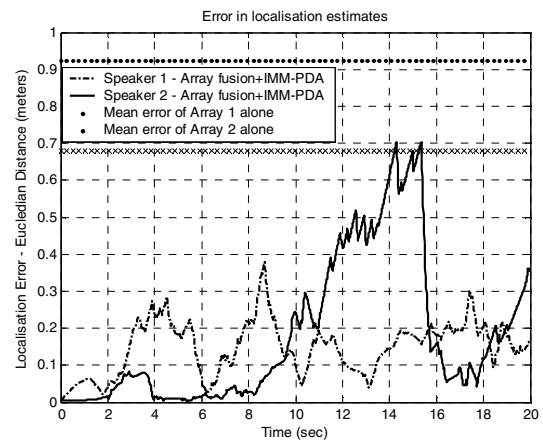


Fig. 8: Absolute localization error for all array configurations.

5. References

- [1] Bar-Shalom Y., Li X., Kirubarajan T., "Estimation with application to tracking and navigation", Wiley, 2001.
- [2] Blackman S., Popoli R., "Design and analysis of modern tracking systems", Artech House, 1999.
- [3] Chen H., et al., "Multiple Target Tracking with Multiple Finite Resolution Sensors", 5th International Conference on Information Fusion, 2002.
- [4] Aarabi P., Zaky S., "Robust sound localization using multi-source audiovisual information fusion", Elsevier, Information Fusion, 2, pp. 209-223, 2001.
- [5] Beal M., Attias H., Jojic N., "Audio visual sensor fusion with probabilistic Graphical models", 7th European Conference on Computer Vision, 1, pp. 736-750.
- [6] Brandstein M., "A Framework for Speech Source Localization Using Sensor Arrays", PhD thesis, Brown University, Providence, RI, 1995.
- [7] Willner D., et al., "Kalman filter algorithms for a multi sensor system", IEEE Conf. on Decision & Control, 1976.
- [8] Sturim D., Brandstein M., Silverman H., "Tracking Multiple Talkers using Microphone Array Measurements", IEEE Proc. ICASSP, pp. 371-374, 1997.
- [9] Mazar E., Averbuch A., Bar-Shalom Y., Dayan J., "IMM methods in target tracking", IEEE Trans. on Aerospace and Electronics Systems, vol 34, no.1, pp. 103-123, 1998.
- [10] Johnson D., Dudgeon D., "Array Signal Processing: Concepts and Techniques", Prentice Hall, 1993.